Eur. Phys. J. Plus (2013) **128**: 117
DOI 10.1140/epjp/i2013-13117-x

THE EUROPEAN PHYSICAL JOURNAL PLUS

Regular Article

Search for the Standard Model Higgs boson in the $H \longrightarrow ZZ \longrightarrow \ell^+\ell^-q\bar{q}$ decay channel at CMS on 4.6 fb^{-1} of 7 TeV proton-proton collision data*

Francesco Pandolfi[a]

ETH Zurich, Switzerland

Received: 21 May 2013 / Revised: 31 July 2013
Published online: 17 October 2013 –

Abstract. A search for the standard model Higgs boson decaying to two Z bosons with subsequent decay to a final state containing two leptons and two quark-jets, $H \rightarrow ZZ \rightarrow \ell^+\ell^-q\bar{q}$ is presented. Results are based on data corresponding to an integrated luminosity of 4.6 fb^{-1} of proton-proton collisions at $\sqrt{s} = 7$ TeV and collected with the CMS detector at the CERN LHC. The analysis performance is strengthened against jet resolution effects thanks to an accurate jet energy calibration carried out on photon+jet events, and the use of a kinematic fit to the $Z \rightarrow q\bar{q}$ decay chain. Selections to discriminate between signal and background events are based on kinematic and topological quantities including the angular spin correlations of the decay products. Events are further classified for analysis according to the probability of the jets to originate from quarks of light or heavy flavor or from gluons. No evidence for a Higgs boson is found and upper limits on the Higgs boson production cross section are set in the range of masses between 200 GeV and 600 GeV.

1 Introduction

Even if, throughout decades of experiments, the Standard Model of elementary particles has proved to be one of the most successful scientifical theories ever elaborated, it can describe a universe populated by massive particles only if its Lagrangian's symmetry is broken. The simplest way to accomplishes this is known as the Higgs mechanism [1–4], which postulated the existence of a scalar boson, the Higgs particle, which, before the contruction of CERN's Large Hadron Collider, had never been observed.

The Large Hadron Collider (LHC) [5] is the particle accelerator which has been built with the aim of producing definitive proof regarding the Higgs boson's existence. It is a superconducting proton collider, which has delivered collisions at center-of-mass energies of 7 and 8 TeV.

The Compact Muon Solenoid (CMS) [6] is one of the four main experiments which analyses the collisions produced at the Large Hadron Collider. It is a general-purpose detector which has been designed to maximise its performance in Higgs boson searches. In July 2012, together with the ATLAS [7] experiment, it has announced the discovery of a new resonance [8,9] which has characteristics compatible with a Higgs boson with mass $m_H \sim 125$ GeV.

This article describes the search for a heavy ($m_H \geq 200$ GeV) Higgs boson in the $H \rightarrow ZZ \rightarrow \ell^+\ell^-q\bar{q}$ decay channel, conducted at CMS on a total of 4.6 fb^{-1} of 7 TeV collision data recorded in 2011. It is organized as follows: sect. 2 describes the experimental setup and defines the physics objects which are to be used at analysis level; sect. 3 will focus on jet energy calibration and performance; sect. 4 will introduce a discriminant capable of discerning between quark and gluon jets; the analysis strategy and the event selection definition is outlined in sect. 5; sect. 6 will detail the background estimation procedure; possible sources of systematic uncertainties are listed in sect. 7; and the results of the analysis, together with their statistical interpretation, are presented in sect. 8. This article summarizes the work of a Ph.D. thesis which constituted the basis of the published result of the high-mass $H \rightarrow ZZ \rightarrow \ell^+\ell^-q\bar{q}$ search performed at CMS [10].

* This paper is based on the author's phD thesis, that was awarded the INFN "Marcello Conversi" prize in 2012.

[a] e-mail: francesco.pandolf@cern.ch

2 Experimental setup and object definitions

A detailed description of the CMS detector can be found elsewhere [11]. Its central feature is a 3.8 T superconducting solenoid of 6 m internal diameter. Within its field volume are the silicon tracker, the crystal electromagnetic calorimeter (ECAL), and the brass/scintillator sampling hadron calorimeter (HCAL). The muon system, composed of drift tubes, cathode strip chambers, and resistive-plate chambers, is installed outside the solenoid, embedded in the steel return yoke. CMS uses a right-handed coordinate system, with the origin at the nominal interaction point, the x-axis pointing to the center of the LHC, the y-axis pointing up (perpendicular to the LHC plane), and the z-axis along the counterclockwise-beam direction. The polar angle θ is measured from the positive z axis and the azimuthal angle ϕ is measured in the x-y plane. The pseudorapidity η is defined as $-\ln[\tan(\theta/2)]$.

Muons [12] are measured with the combination of the tracker and the muon system, in the pseudorapidity range $|\eta| < 2.4$. Electrons [13] are detected as tracks in the tracker pointing to energy clusters in the ECAL up to $|\eta| = 2.5$. The full details of electron and muon identification criteria are described elsewhere [14]. Isolation requirements on lepton candidates are enforced by measuring the additional detector activity in a surrounding cone of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.3$, where $\Delta\eta$ and $\Delta\phi$ are the differences in pseudorapidity and in azimuthal angle, measured in radians, respectively. For muons the total scalar sum of the transverse momenta of the additional reconstructed tracks and of the energy in the calorimeters in the surrounding cone is required to be less than 15% of the muon transverse momentum. Electron isolation requirements are similar but vary depending on the shape of the reconstructed energy distribution in the electromagnetic calorimeter.

Photons are reconstructed from ECAL energy deposits, and identified with isolation and cluster shape criteria. Isolation requirements are enforced both in the tracker and in the calorimeters: the scalar sum of reconstructed track transverse momenta within $\Delta R = 0.35$ about the photon candidate direction is required not to exceed 10% of the photon p_{T}, and the total calorimetric energy within $\Delta R = 0.4$, excluding the energy associated to the photon, must be less than 5% of the photon energy. Cluster shape criteria are enforced through the second moments of the energy distribution of the photon seed basic cluster in the direction of the cluster principal axes: the deposit is required to be compatible with an electromagnetic shower produced by a single energetic photon.

Jets are reconstructed with a Particle-Flow (PF) algorithm [15], a global event reconstruction technique which optimally combines the information of all sub-detectors to reconstruct the particles produced in a collision. Reconstructed particle candidates (PFCandidates) are clustered to form PF jets with the *anti-*k_{T} [16] with a distance parameter of 0.5. The jet energy resolution is typically 15% at 10 GeV and 8% at 100 GeV. Jets are required to be inside the tracker acceptance ($|\eta| < 2.4$), to increase the reconstruction efficiency and the precision of the energy measurement using PF techniques. Jet energy corrections are applied to account for the non-linear response of the calorimeters to the particle energies and other instrumental effects. These corrections are based on in situ measurements using dijet and γ+jet data samples [17]. Simultaneous proton-proton collisions within the same bunch crossing (pile-up) has an effect on jet reconstruction by contributing additional particles to the reconstructed jets. The average energy density due to pile-up (ρ_{PF}) is evaluated in each event and the corresponding energy is subtracted from each jet [18]. A jet identification requirement, primarily based on the energy balance between charged and neutral hadrons in a jet, is applied to remove misidentified jets. Jets are required to have $p_{\mathrm{T}} > 30$ GeV.

To identify jets originating from the hadronization of bottom quarks, a b-tagging algorithm [19] is employed. The algorithm identifies jets from b-hadron decays by requiring at least two tracks to have significant impact parameters with respect to the primary interaction vertex. This tagger is used here with two operating points: the *loose* point corresponds to an efficiency for jets originating from bottom quarks of about 80% and a misidentification probability for jets from light quarks and gluons of 10%, while the *medium* operating point provides an efficiency for b-jets of about 65% and a misidentification probability of about 1%.

Missing transverse energy is also defined with the particle-flow approach, as the norm of the vectorial sum of the transverse momenta of all PFCandidates reconstructed in the event.

This measurement uses data from proton-proton collisions, produced at a center-of-mass energy of 7 TeV, corresponding to an integrated luminosity of 4.6 fb^{-1}. The data were collected by the Compact Muon Solenoid detector at the Large Hadron Collider in 2011. The data were recorded through the dilepton high level trigger paths, which require the presence, in the event, of a pair of hard reconstructed muons or electrons. To limit the contamination of fake electron candidates reconstructed within jets, isolation requirements are employed in the di-electron triggers. The muonic dataset, finally, has been increased by about 6% by adding to the events triggered by the dimuon triggers those recorded through isolated single-muon paths.

Although the analysis makes use of a completely data-driven background estimation procedure, simulated backgrounds were analyzed to optimize the event selection criteria. The main backgrounds, as will be seen, are constituted by the production of a leptonically-decaying Z boson in association with hard jets, and the production of top-antitop pairs. Both these processes have been generated with MADGRAPH 4.4.12 [20], a leading order matrix element generator, interfaced with PYTHIA 6.4 [21] for parton showering and hadronization. Other backgrounds have been completely generated in PYTHIA 6.4. Signal events have instead been generated with the POWHEG [22–24] box, which contains the complete NLO calculation of the process, which is also interfaced to PYTHIA. Generated events

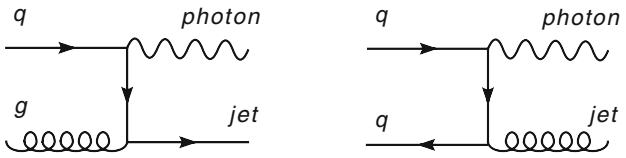

Fig. 1. Dominant photon+jet production diagrams at a proton-proton collider.

are then passed through a full simulation of the CMS detector, implemented in the GEANT 4 [25] software framework. Multiple minimum-bias events are superimposed to the hard scattering event to simulate pile-up. All simulated events have been reweighted to take into account the pile-up profile of the data, where a mode of seven interactions per bunch crossing was observed. No explicit trigger requirement was made on Monte Carlo events, but these were rescaled in order to take into account the measured trigger efficiencies in data.

3 Jet reconstruction performance in photon+jet events

Two variables are employed to measure jet reconstruction performance: the jet response and resolution. The response is defined on a jet-by-jet basis by matching, in the simulation, each reconstructed jet to the corresponding generator jet, which is defined by applying the same clustering algorithm to stable generator particles produced during hadronization. It is hence defined as the ratio between the transverse momenta of the reconstructed jet ($p_{\mathrm{T}}^{\mathrm{reco}}$) and the generator jet ($p_{\mathrm{T}}^{\mathrm{gen}}$):

$$R = \frac{p_{\mathrm{T}}^{\mathrm{reco}}}{p_{\mathrm{T}}^{\mathrm{gen}}} \,. \tag{1}$$

Its average value, $\langle R \rangle$, is an estimator of the response of a given jet reconstruction strategy. We will call this the *true* response in the following. The jet *true resolution*, instead, is defined as the RMS of the R variable distribution, divided by the true response.

The strategy adopted by CMS is to derive jet energy corrections from the full-detector software simulation of jets and apply them on the data. The data are then used to test their effectiveness, and if a significant non-closure is found, an additional (residual) correction is introduced. This approach allows to minimize the effect of statistical fluctuations deriving from insufficient events in the data samples.

The measurement of the absolute jet energy scale is done with photon+jet events, with a technique first introduced at Tevatron experiments [26]. Their dominant production diagrams at a proton-proton collider are shown in fig. 1. At leading order, in these events the photon and the leading jet are balanced in the transverse plane, hence the precision with which the photon is measured in the crystal ECAL ($\sim 1\%$) can be exploited to infer the true jet transverse momentum, and therefore measure the reconstructed jet response.

The measurement presented in this article makes use of the first $1\,\mathrm{fb}^{-1}$ of data recorded by the CMS detector during the 2011 data taking through the single photon HLT paths, which require the presence of a high-transverse-momentum energy deposit in the ECAL. The event selection requires the photon candidate to be in the ECAL barrel fiducial region ($|\eta| < 1.3$), and to have transverse momentum greater than 15 GeV. The data are compared to simulated photon+jet events generated with PYTHIA 6. The expected sample purity after these requirements is expected to be of the order of 90% for photon transverse momenta greater than 100 GeV, and somewhat worse for lower transverse momenta. The dominant background is constituted by QCD dijet events, in which one jet is misidentified as a photon. The bias introduced by this background is expected to play a minor role: QCD events which pass the selection will present a parton which has hadronized mainly into one (or more) electromagnetic-decaying particles, so these events are very similar to true photon+jet events for practical purposes.

3.1 Photon-jet balancing

We define the reconstructed balancing response estimate as the ratio between the jet and the photon transverse momenta:

$$R_{\mathrm{balancing}} = \frac{p_{\mathrm{T}}^{\mathrm{recoJet}}}{p_{\mathrm{T}}^{\gamma}} \,.$$

It is always possibile to factorize it in the following manner:

$$R_{\mathrm{balancing}} = \frac{p_{\mathrm{T}}^{\mathrm{recoJet}}}{p_{\mathrm{T}}^{\gamma}} = \frac{p_{\mathrm{T}}^{\mathrm{recoJet}}}{p_{\mathrm{T}}^{\mathrm{genJet}}} \cdot \frac{p_{\mathrm{T}}^{\mathrm{genJet}}}{p_{\mathrm{T}}^{\gamma}} \,, \tag{2}$$

where we have introduced the transverse momentum of the generator jet matched to the reconstructed jet.

The new expression presents two factors. By comparing to eq. (1) one can easily recognize the true response variable in the first ratio. We will define this ratio as the *intrinsic* response, and it depends on the chosen jet reconstruction scheme and on the jet transverse momentum. It is the object of the jet energy scale measurement.

The second ratio, on the other hand,

$$\frac{p_{\mathrm{T}}^{\mathrm{genJet}}}{p_{\mathrm{T}}^{\gamma}},$$

is a measure of the imbalance at generator level between the photon and the leading jet. It depends on the amount of additional event activity, and on the efficiency of the chosen jet algorithm. We will call it generically *imbalance*.

Imbalance is the main source of bias in estimating the jet energy scale with photon+jet balancing. In order to reduce its effects a requirement on the transverse momentum of the subleading jet is introduced:

$$p_{\mathrm{T}}^{\mathrm{2ndJet}} < \max\left(0.1 \cdot p_{\mathrm{T}}^{\gamma},\ 5\,\mathrm{GeV}\right). \tag{3}$$

This requirement, though, does not eliminate all of the bias. In order to do so, more sophisticated approaches are needed. Two methods have been devised at CMS to minimize the bias originating from imbalance: the Missing-E_{T} Projection Fraction, and the balancing extrapolation.

3.1.1 Missing-E_{T} Projection Fraction method

The Missing-E_{T} Projection Fraction (MPF) method was first employed at the D0 detector [26], and, as it makes use of the event reconstruction as a whole, turns out to be particularly well suited for particle-flow reconstruction. It stems from the basic assumption that at generator level the vectorial sum of the transverse momenta of all final state objects must cancel out on a per-event basis. In photon+jet events we can group these objects in two groups: the photon, and the rest of the event, which we will call the hadronic recoil. Therefore for each event it holds

$$\boldsymbol{p}_{\mathrm{T}}^{\gamma,\mathrm{MC}} + \boldsymbol{p}_{\mathrm{T}}^{\mathrm{recoil}} = \mathbf{0}.$$

When folding in the detector finite responses and resolutions, we obtain

$$R_{\gamma}\boldsymbol{p}_{\mathrm{T}}^{\gamma,\mathrm{MC}} + R_{\mathrm{recoil}}\,\boldsymbol{p}_{\mathrm{T}}^{\mathrm{recoil}} = -\boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}},$$

where R_{γ} and R_{recoil} denote, respectively, the detector response to the photon and the recoil, and $\boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}}$ is the event missing transverse energy. Solving for $R_{\mathrm{recoil}}/R_{\gamma}$ and defining $\boldsymbol{p}_{\mathrm{T}}^{\gamma,\mathrm{reco}} \equiv R_{\gamma}\boldsymbol{p}_{\mathrm{T}}^{\gamma,\mathrm{MC}}$ yields

$$R_{\mathrm{recoil}}/R_{\gamma} = 1 + \frac{\boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}} \cdot \boldsymbol{p}_{\mathrm{T}}^{\gamma,\mathrm{reco}}}{|\boldsymbol{p}_{\mathrm{T}}^{\gamma,\mathrm{reco}}|^2} \equiv R_{\mathrm{MPF}},$$

which defines the MPF response variable.

As it considers the hadronic recoil as a whole, the MPF response variable proves to be robust, showing very low sensitivity to additional event activity and pile-up. It further is an unbiased estimator of the jet response, as long as most of the recoil energy is carried by the leading jet in the event. This condition is fulfilled with a simple cut on the subleading jet transverse momentum, such as the one presented in eq. (3).

In each photon transverse momentum bin the response estimate is derived with a 99%-truncated mean, both with the balancing and the MPF variables. The result, as a function of the photon p_{T}, is shown in fig. 2: the left graph shows the trend of the response estimates, for data and MC (markers), and compares them to the true response (black line); data-MC ratios are shown in the right graph. As can be seen, the simple balancing estimate presents a visible bias in the measurement of the jet response for $p_{\mathrm{T}} < 80\,\mathrm{GeV}$.

3.1.2 Balancing extrapolation method

The second method which minimizes the imbalance bias is the balancing extrapolation. This method is still based on a simple balancing between the leading jet and the photon, but instead of reducing the effect of additional event activity by imposing a requirement on the subleading jet, it studies the trend of the response as a function of the subleading jet's transverse momentum. The trend is then extrapolated to the ideal case of no secondary jet activity, with photon and leading jet perfectly balanced in the transverse plane. Differently from what was presented in the previous section, in which the MPF method was used to measure the uncorrected jet response, we here will show the response of PFJets which have undergone the full set of CMS jet energy corrections, as a verification of the validity of such corrections.

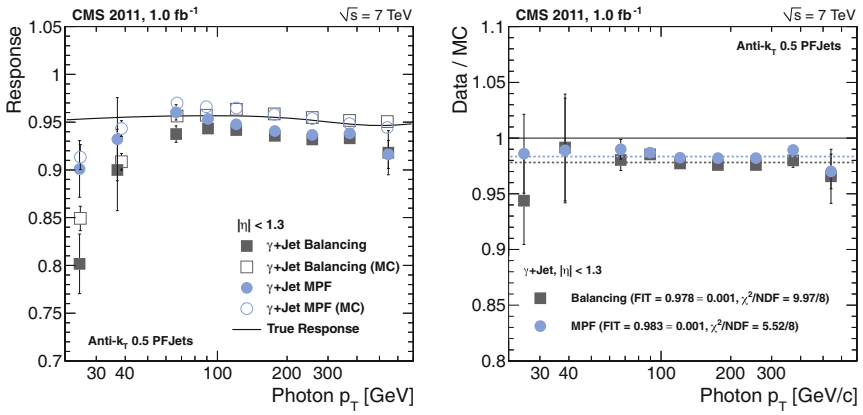

Fig. 2. Measurement of the response of *anti-*k_{T} 0.5 PFJets in the CMS barrel ($|\eta| < 1.3$). Left: response as a function of photon transverse momentum for the balancing (grey squares) and MPF (blue circles) methods, in 1.0 fb^{-1} of data (solid) and in the MC simulation (hollow). A comparison to the true response (black line) is also shown. Right: data/MC ratios.

In a given photon transverse momentum range, recalling eq. (2), which we may rewrite as $R_{\mathrm{balancing}} = R_{\mathrm{intr}} \cdot R_{\mathrm{imb}}$, we expect:

- the intrinsic response R_{intr} to be independent of the subleading jet (as long as it is "reasonably" small), as it concerns only the leading jet;
- the imbalance R_{imb} to have a strong dependance on the subleading jet.

Our assumption is that these two effects are not correlated, so that they factorize, and therefore the response and resolution will have simple expressions:

$$\begin{aligned} \langle R_{\mathrm{balancing}} \rangle &= \langle R_{\mathrm{intr}} \rangle \cdot \langle R_{\mathrm{imb}} \rangle \\ \sigma_{\mathrm{balancing}} &= \sigma_{\mathrm{intr}} \oplus \sigma_{\mathrm{imb}} \end{aligned} \tag{4}$$

where we have used the symbols $\langle R \rangle$ and σ to indicate respectively response and resolution.

For what concerns the response, empirically we find that the functional dependance of R_{imb} on the subleading jet p_{T} is of quadratic form. Therefore, in a given photon p_{T} bin we will have

$$\begin{aligned} \langle R_{\mathrm{intr}} \rangle (p_{\mathrm{T}}^{\mathrm{2ndJet}}) = c \qquad \langle R_{\mathrm{imb}} \rangle (p_{\mathrm{T}}^{\mathrm{2ndJet}}) = 1 - q - m(p_{\mathrm{T}}^{\mathrm{2ndJet}})^2 \qquad c, q, m = \mathrm{const} \\ \Rightarrow \langle R_{\mathrm{balancing}} \rangle (p_{\mathrm{T}}^{\mathrm{2ndJet}}) = c \cdot \left[1 - q - m(p_{\mathrm{T}}^{\mathrm{2ndJet}})^2 \right], \end{aligned} \tag{5}$$

therefore c is the object of this measurement, m describes the dependance of the imbalance on the subleading jet, and q quantifies the amount of irreducible imbalance between the photon and the leading jet. The values assumed by q in the simulation are negative and as large as -5% at low transverse momenta (dominated by jet algorithm inefficiencies), positive and of the order of $+1\%$ at very high transverse momenta (dominated by photon energy scale effects).

The method's operation is shown in fig. 3 (left), where the trends of the different contributions are shown as a function of the relative subleading jet transverse momentum ($p_{\mathrm{T}}^{\mathrm{2ndJet}}/p_{\mathrm{T}}^{\gamma}$), for events with photons with transverse momentum between 100 and 150 GeV. The intrinsic response (blue squares) and the imbalance (black triangles) can be seen, together with their fit functions. The product of these two functions is shown with a grey line, and, if the made assumptions are correct, should constitute the predicted trend for the pseudo data points (open red markers). The observed good agreement between the two is a confirmation of the validity of the method on the simulation.

The measured trends in the data are also shown in each graph with solid red markers. The effect of the irreducible imbalance cannot be measured on data but must be accounted for, therefore the function used in the fit to the data has the functional form defined in eq. (5), but with the q parameter fixed to the value obtained on the simulation.

The measured corrected response as a function of photon transverse momentum are also shown in fig. 3. The center plot shows the extrapolated response values, in the data and in the simulation, for simple balancing (grey) and for the extrapolation method (red), together with the expected true response (black line). The latter is visibly larger than unity at low transverse momenta: this is caused by the fact that the jet energy corrections are derived on QCD events, which are dominated by gluon jets, which have lower response than quark jets, that dominate the photon+jet events studied in this analysis. The right plot in fig. 3 shows the data-MC ratios of the two methods. Consistently with what found with the MPF method on uncorrected response in the previous section, the data present a response about 1.5% lower than what the simulation predicts. This constitutes the *residual* absolute jet energy scale correction, which is applied to the data after the MC-based corrections and the residual relative scale corrections.

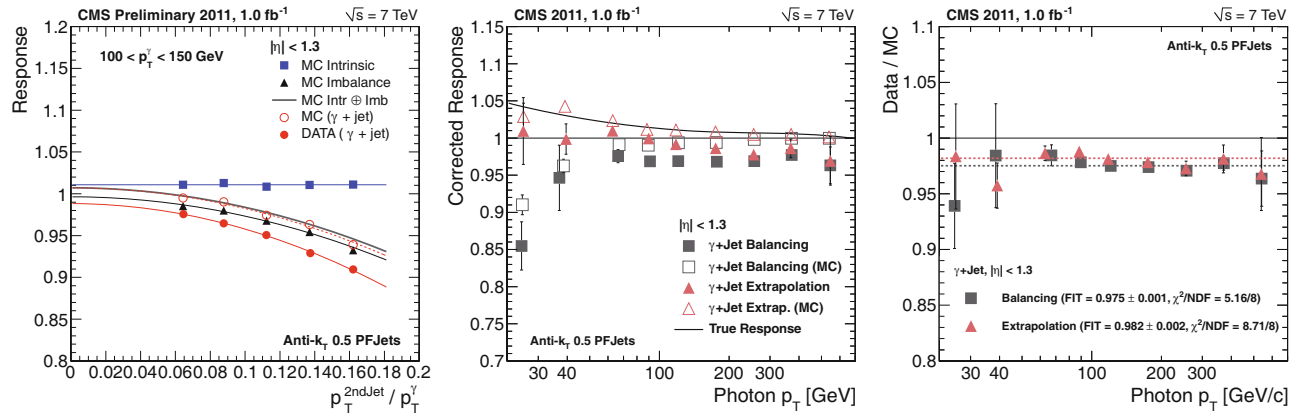

Fig. 3. Right: balancing response extrapolation in a representative transverse-momentum range, for *anti-k_T* 0.5 PFJets reconstructed in the barrel. Center: corrected response measurement, as a function of photon transverse momentum, for *anti-k_T* 0.5 PFJets reconstructed in the barrel. Results for balancing (grey) and extrapolation (red) are shown both for data (solid) and the Monte Carlo simulation (hollow). A comparison to the expected true response (black line) is also shown. Right: data-MC ratios for the two methods.

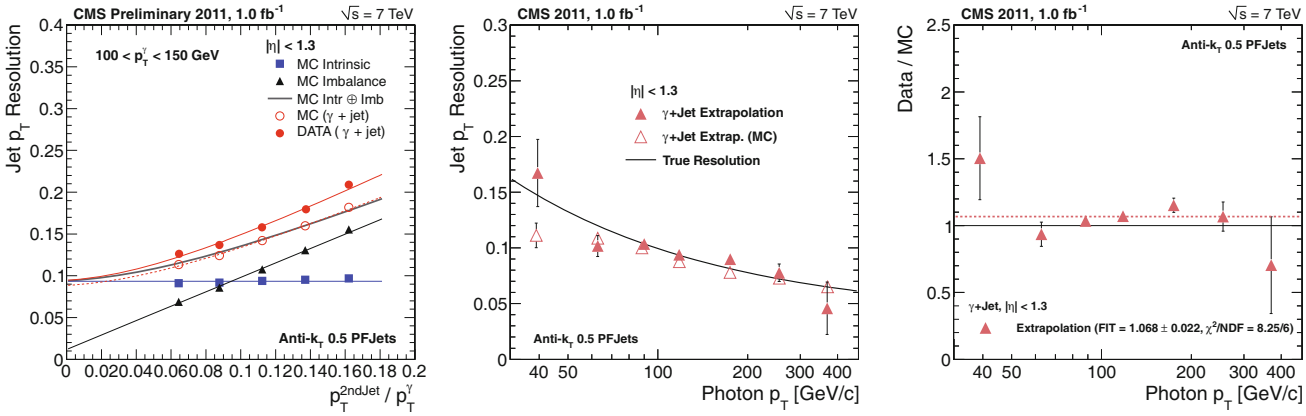

Fig. 4. Left: balancing resolution extrapolation in a representative transverse-momentum range, for *anti-k_T* 0.5 PFJets reconstructed in the barrel. Center: jet p_T resolution measurement, as a function of photon transverse momentum, for *anti-k_T* 0.5 PFJets reconstructed in the barrel, in data and MC. Right: data-MC ratio.

3.2 Jet transverse-momentum resolution measurement

The balancing extrapolation method allows us to measure also the corrected jet transverse-momentum resolution. Recalling eq. (4), our assumptions are that, in a given p_T^γ bin, the intrinsic resolution is independent of p_T^{2ndJet}, whereas the imbalance effect to be linear. In formulas

$$\sigma_{intr}(p_T^{2ndJet}) = c' \qquad \sigma_{imb}(p_T^{2ndJet}) = q' + m' \cdot p_T^{2ndJet} \qquad c', q', m' = \text{const}$$
$$\Longrightarrow \sigma_{balancing}(p_T^{2ndJet}) = \sqrt{c'^2 + q'^2 + 2q'm' \cdot p_T^{2ndJet} + m'^2 \cdot \left(p_T^{2ndJet}\right)^2}\,.$$

An example of the performance of the method is shown in fig. 4 (left), for events with photon transverse momentum between 100 and 150 GeV, in the data and in the simulation. The colour coding is the same as in the response case. Again, the good agreement between the "predicted" trend (grey line) and the reconstructed MC estimates (open red circles) proves the internal consistency of the method. The data points are fitted with the expected functional form, and, similarly as in the response case, the contribution of the irreducible imbalance (q') is fixed to the value fitted in the MC.

The results of the corrected jet p_T resolution as a function of transverse momentum are also shown in fig. 4. The center plot shows the results of the extrapolation, in data and MC, and compares them to the true resolution (black line). The right plot shows the ratio of the measurements in data and MC: the resolution measured in the data are found to be about 7% worse than the MC.

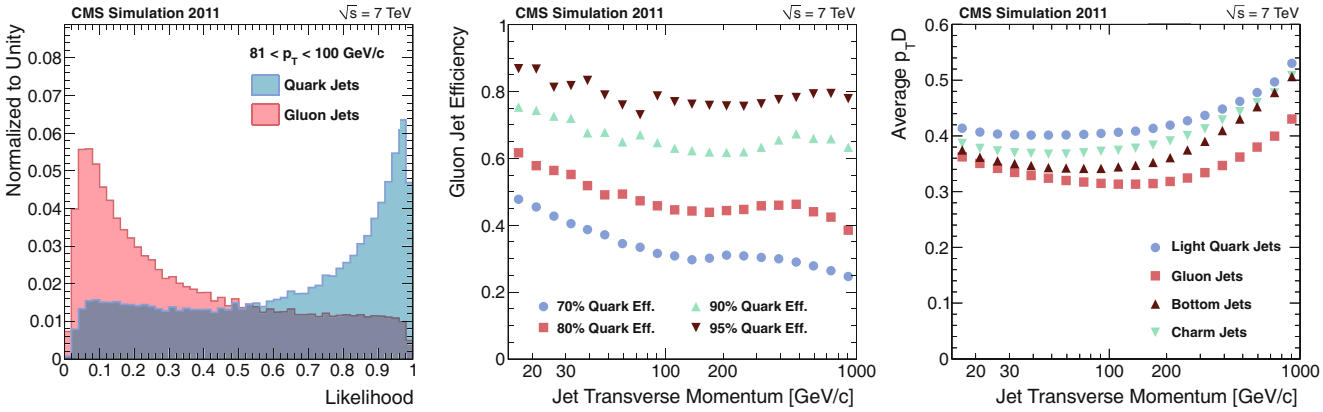

Fig. 5. Left: probability density distributions of the quark-gluon likelihood discriminant for quark (blue) and gluon (red) jets with transverse momentum between 81 and 100 GeV. Center: gluon jet efficiency, as a function of transverse momentum, for fixed values of quark jet efficiency: 70% (blue circles), 80% (red squares), 90% (green upwards triangles) and 95% (downwards brown triangles). Right: average $p_{\mathrm{T}}D$ as a function of jet transverse momentum for light quarks (blue circles), gluons (red squares), bottom quarks (brown upwards triangles) and charm quarks (green upwards triangles).

4 Quark-gluon discrimination

Detailed information on jet composition and substructure, as the one provided by the particle-flow reconstruction, may be exploited to gain insight on the nature of the jet's underlying parton. Gluons and quarks have different colour interaction, and this will mirror in their hadronization: gluons will favor wider, high-multiplicity jets, when compared to those generated by final state (light) quarks. Furthermore, the phenomenon of "gluon-splitting", if occurring at the beginning of hadronization, may give rise to jets made of a number of collimated quark sub-jets.

These structural differences between gluon and quark hadronization may be exploited to derive a likelihood-based discriminant. In order to do so, the most precise and granular information on the jet particle composition must be accessed, such as the one provided by the CMS particle-flow event reconstruction.

We have studied the use of three variables:

- *charged hadron multiplicity*: the number of charged hadron PFCandidates clustered in the jet;
- *neutral multiplicity*: the number of PFCandidates in the jet which are photons or neutral hadrons;
- *transverse momentum distribution* ($p_{\mathrm{T}}D$) among PFCandidates inside the jet, defined as:

$$p_{\mathrm{T}}D = \sqrt{\frac{\sum p_{\mathrm{T}}^2}{(\sum p_{\mathrm{T}})^2}},$$

where the sums are extended to all PFCandidates inside the jet. It stems from its definition that $p_{\mathrm{T}}D \to 1$ for a jet made of one single candidate which carries the totality of its momentum, whereas $p_{\mathrm{T}}D \to 0$ for jets composed of an infinite number of particles.

Probability density functions (PDFs) are defined on simulated QCD events for these three variables, separately for jets which are originated from light quark and gluon jets. These PDFs are then combined into a likelihood discriminant, taken as a simple product of the three variables. To take into account the fact that the variables depend strongly both on the jet transverse momentum and on the amount of pile-up activity of the event, the phase space is subdivided with a two-dimensional binning: 20 transverse-momentum bins from 15 to 1000 GeV are multiplied by 17 intervals in the particle-flow event energy density variable ρ_{PF}, from 0 to 17 GeV.

Figure 5 (left) shows the distributions of the likelihood estimator variable for light quark and gluon jets with transverse momentum between 81 and 100 GeV. Figure 5 (center) shows instead the likelihood's discriminating performance, in terms of maximum achievable gluon jet rejection as a function of the jet transverse momentum for four different light quark jet efficiency working points (70%, 80%, 90% and 95%). As can be seen, the discriminating performance of the estimator is worst at low transverse momenta, gradually improves up to about 100 GeV, where it reaches a plateau which is maintained up to the TeV scale.

As has been previously observed by LEP experiments [27], the hadronization of a bottom quark yields jets which have structures similar to gluon-initiated jets, from an experimental point of view. This can be seen in fig. 5 (right), where the average of the $p_{\mathrm{T}}D$ variable is shown as a function of jet transverse momentum, for light quarks (blue circles), gluons (red squares), bottom quarks (brown upwards triangles) and charm quarks (green upwards triangles): up to

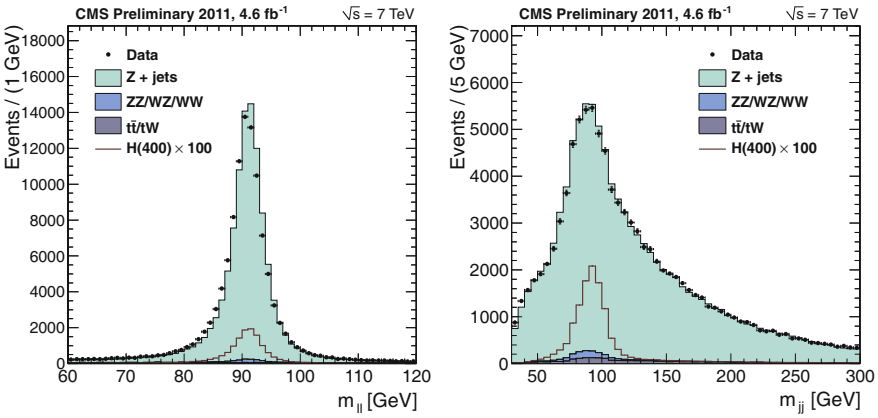

Fig. 6. Dilepton (left) and dijet (right) invariant mass distributions. Events passing preselection in $4.6\,\mathrm{fb}^{-1}$ of 2011 data are compared to the expected yield of the dominant backgrounds in the simulation. The distribution for events coming from a 400 GeV Higgs boson decay, enhanced by a factor 100, is superimposed.

transverse momenta of about 100 GeV, bottom-quark initiated jets tend to be more similar to gluon jets than to light quark jets. The other two variables are found to present similar trends. We therefore conclude that this discriminator is not effective in discriminating bottom quarks from gluons, and hence, in rejecting gluon jets while keeping quark jets, must be used on a sufficiently bottom-deprived jet sample.

5 Event selection

As this analysis searches for a heavy Higgs boson, the signature of signal events presents two energetic Z bosons, one decaying to a pair of electrons or muons, the other to jets. Therefore the event preselection is defined as those events which contain two oppositely charged electrons or muons with transverse momenta respectively greater than 40 and 20 GeV, and two or more jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$. The relatively high transverse momentum requirement on the lepton pair is introduced to ensure high trigger efficiency. In the case of multiple electron or muon pairs, the oppositely charged pair with invariant mass closest to the Z boson nominal mass is chosen. Events are then correctly identified as signal event candidates if the invariant mass of the dilepton system lies between 70 and 110 GeV. If an event is found to present both an electron and a muon pair passing this requirement, it is discarded. The dilepton invariant mass for events passing preselection requirements is shown in fig. 6 (left).

The event is further required to present at least one jet pair with an invariant mass in the 75–105 GeV range. The requirement on the hadronic invariant mass is more stringent than the leptonic one, for it is a powerful handle in discriminating the main backgrounds, which do not present a real Z boson decaying to jets. It is therefore kept closest to the nominal Z boson mass, compatibly with the expected dijet mass resolution, which is about 15 GeV for signal events. Events which pass the dilepton mass requirement but not the dijet one are nevertheless kept, and are categorized as sideband events, as will be explained in sect. 6.

In general, though, a signal event candidate will present multiple jet pairs. This is true also for true signal events, as additional jets will be created in proton fragmentation or in the process of creation of the Higgs boson. In order to minimize the effect of signal self-combinatorics, the jet pair with the invariant mass closest to the Z boson nominal mass will be selected, even if in the context of the event categorization procedure which will be described in detail in sect. 5.1. The distribution of the invariant mass of the dijet pair with mass closest to the Z mass for events which pass preselection requirements is shown in fig. 6 (right).

5.1 Categorization

A cardinal point of this analysis is understanding that jet flavour may provide a powerful means of background discrimination. From a jet flavour point of view the main differences between signal and background jets are the relatively large contribution of heavy flavour quarks (b and c) and the absence of gluons. We take advantage of both features in the analysis by pursuing two directives: isolate heavy flavours, in order to identify an event sub-population in which only a fraction of the signal is present, but with a higher expected purity, as backgrounds are less present; limit the background gluon infiltration, trying to affect signal efficiency in a minor way.

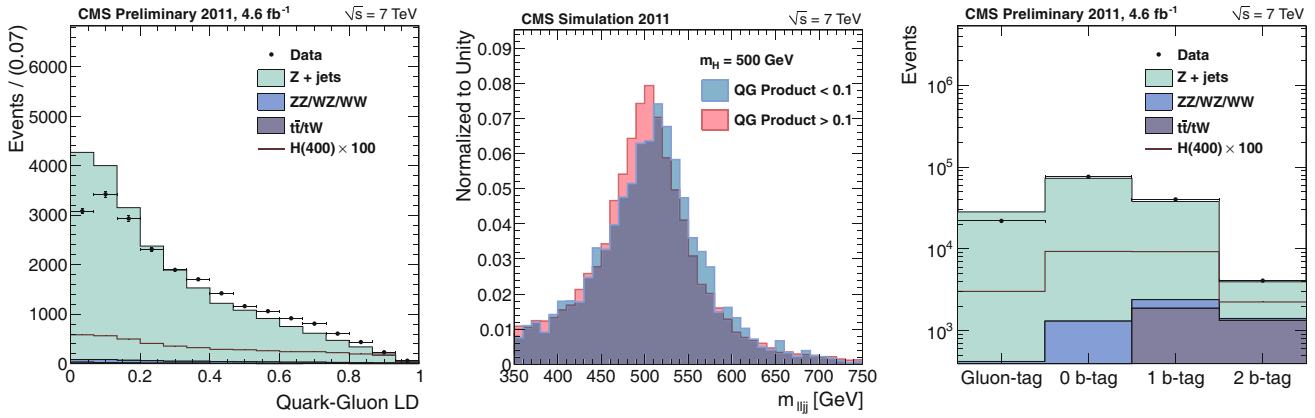

Fig. 7. Left: quark-gluon likelihood product for events passing preselection in 4.6 fb^{-1} of 2011 data and the expected yield of the dominant backgrounds in the simulation; the distribution for events coming from a 400 GeV Higgs boson decay, enhanced by a factor 100, is superimposed. Center: reconstructed Higgs candidate invariant mass distribution for events passing (red) and failing (blue) the Q-G likelihood requirement, for signal events with a hypothetical Higgs boson masses of 500 GeV; distributions are normalized to unit area. Right: distribution of flavour tagging categories in events passing the analysis preselection in 4.6 fb^{-1} of 2011 data, compared to the expected yield of the dominant backgrounds in the simulation. The distribution for events coming from a 400 GeV Higgs boson decay, enhanced by a factor 100, is superimposed.

In order to identify heavy flavour jets we will use the b-tagging discriminant described in sect. 2, whereas the gluon jet rejection is performed with the likelihood ratio introduced in sect. 4. The analysis will therefore be split into four categories:

- 2 b-tag category: events in which both jets are positively identified as originating from a b quark hadronization;
- 1 b-tag category: events in which one jet is positively identified as a b-jet;
- 0 b-tag category: no jet is identified as b, and the jet pair is not incompatible with a light-quark hypothesis;
- gluon-tag category: events in which jets are likely to originate from gluons.

We expect the 2 b-tag category to have the highest purity, but low signal efficiency, and the 0 b-tag category to have the highest signal efficiency, but large background yields. The gluon-tag category is dominated by background contributions.

An event is placed in the 2 b-tag category if one jet is identified with medium and the other is identified with loose b-tagging requirements. Events which fail these criteria but still contain at least one jet which satisfies the loose criterion are placed in the 1 b-tag category. Events which fail the b-jet identification requirements which would place them in the single- or double-tagged categories, are then split between the 0 b-tag and the gluon-tag categories, by looking at the product of the two jets' quark-gluon (Q-G) likelihood discriminants. Figure 7 (left) shows the distribution of the product of the two jets' Q-G likelihood discriminants, in events passing preselection requirements in data and the simulation. Events with Q-G likelihood product less than 0.1 are rejected and placed in the gluon-tag category. This requirement has an efficiency of about 85% on signal events, and reduces the $Z+\text{jets}$ background by about 34%, 43%, 50%, and 56% at m_{lljj} masses around 250, 300, 400, and 500 GeV.

In addition to being a means of background discrimination, the requirement on the Q-G discriminant also improves the invariant mass resolution in signal events. This is because, by selecting events in which the jet pair has composition properties which are compatible with the expectations for high-p_{T} quark jets, events with misreconstructed jets and events in which signal self-combinatorics leads to the choice of the incorrect jet pair are discarded. This may be seen in fig. 7 (center), where the dilepton-dijet invariant mass for events passing (red) and failing (blue) the requirement on the product of the two jet's Q-G likelihood discriminant, for a signal events with a hypothetical Higgs boson mass of 500 GeV.

In general, an event will have numerous jets, therefore multiple jet pairs. The analysis selection algorithm scans all possible jet pairs, and verifies if the given pair passes the selection requirements, which will depend on the pair's b-tag values, and on the invariant mass of the resulting reconstructed Higgs candidate, as will be described in the following section. If an event presents more than one pair which meets the requirements, the pair which belongs to the highest b-tag category is selected, in order to favor the highest purity samples. If the primacy is shared by more than one pair, the pair with an invariant mass closest to the nominal Z boson mass is selected. This ensures univocal classification of events, and therefore the statistical independence of the samples identified by the categories. Figure 7 (right) shows the subdivision of events in the analysis categories, and as can be seen the background composition can vary significantly among them.

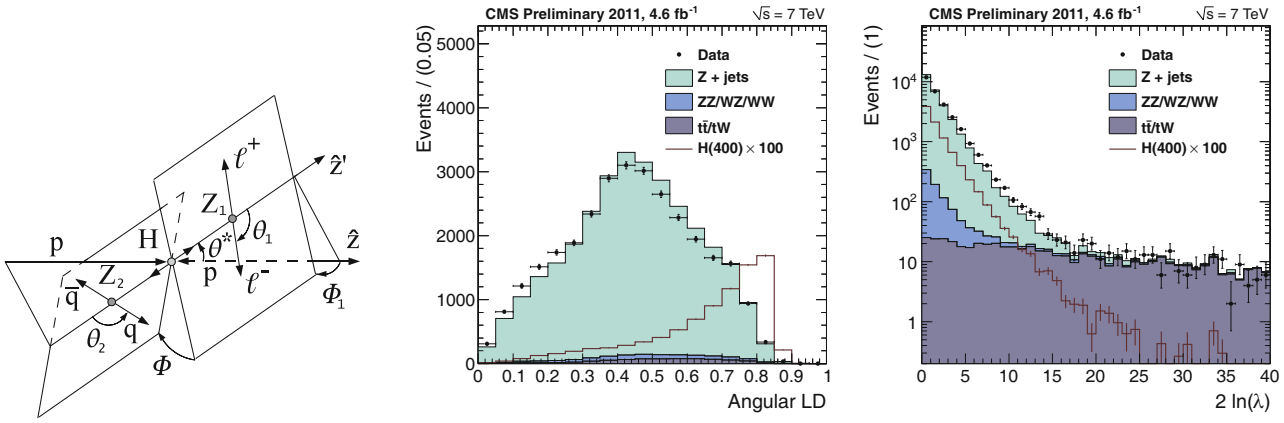

Fig. 8. Right: adopted convention in the definition of the three helicity angles (θ_1, θ_2 and Φ) and two production angles (θ^* and Φ_1) which univocally describe the $H \to ZZ \to \ell^+\ell^-q\bar{q}$ decay chain. Center: data-simulation comparisons for the angular likelihood discriminant. Right: Missing transverse-energy significance ($2\ln\lambda$) distribution in events passing the analysis preselection in $4.6\,fb^{-1}$ of 2011 data, compared to the expected yield of the dominant backgrounds in the simulation. The distributions for events coming from a 400 GeV Higgs boson decay, enhanced by a factor 100, are superimposed.

Table 1. Results of the angular likelihood (aLD) discriminant threshold optimization, in the three b-tag categories.

b-tag category	Optimal aLD threshold
0	$0.55 + 0.00025 \cdot m_H$[GeV]
1	$0.302 + 0.000656 \cdot m_H$ [GeV]
2	0.5

5.2 Angular analysis

In signal events, the spin of the decaying boson defines the correlations between its decay products, as the latter are the product of a very precise decay chain: the decay of a spin-0 boson (the Higgs) to a pair of identical spin-1 bosons (the Z's), which then decay to fermions. In background events, the spin correlation is absent, therefore we expect to observe different final-state angular distributions.

If we do assume that the four final state objects derive from the above mentioned decay chain, the final-state kinematics in the Higgs boson rest frame, once the masses of the secondary particles are fixed, are univocally determined through the definition of five angles. Following the convention used in [28], we will define them as in fig. 8 (left): they are three helicity angles (θ_1, θ_2 and Φ), respectively defined in the $Z \to \ell\ell$, $Z \to jj$ and Higgs boson rest frames, and two production angles (θ^* and Φ_1), both defined in the Higgs rest frame.

The probability density functions for signal events are computed analytically, and corrected with the use of the simulation to take into account acceptance effects. Those for backgrounds events are empirically fitted on the simulation. Once the probability density functions are defined for the five angles, an angular discriminant is constructed as a simple likelihood ratio. The discriminant is defined in such a way that it is defined between 0 and 1, and peaks at high values for signal events, whereas assumes on average lower values for non-resonant backgrounds. Figure 8 (center) shows the observed distributions of the angular likelihood discriminant in the 2011 data collected by CMS, in events which pass the analysis preselection. The data are compared to the summed contribution of all MC backgrounds, and an overall good agreement is observed. The expected distributions for events coming from a 400 GeV Higgs boson decay are superimposed, scaled by a factor 100.

The angular likelihood is the main tool for background discrimination. Selection thresholds have been identified by minimizing, separately in the three b-tag categories, the expected single-category 95% confidence level upper limit on the Standard Model signal. The optimization was carried out at six pivotal hypothetical signal mass points: 250, 300, 350, 400, 450, and 500 GeV. In each b-tag category the trend of the optimal angular likelihood discriminant threshold was studied as a function of the signal mass, and linear dependancies were found. They were therefore fitted with linear functions of the Higgs mass, in order to find a smooth functional dependance on the mass, and the results of this fit are summarized in table 1. It must be noted that, as no significant deviation from a constant threshold was found in the 2-tag category, the requirement of 0.5 was adopted for all masses.

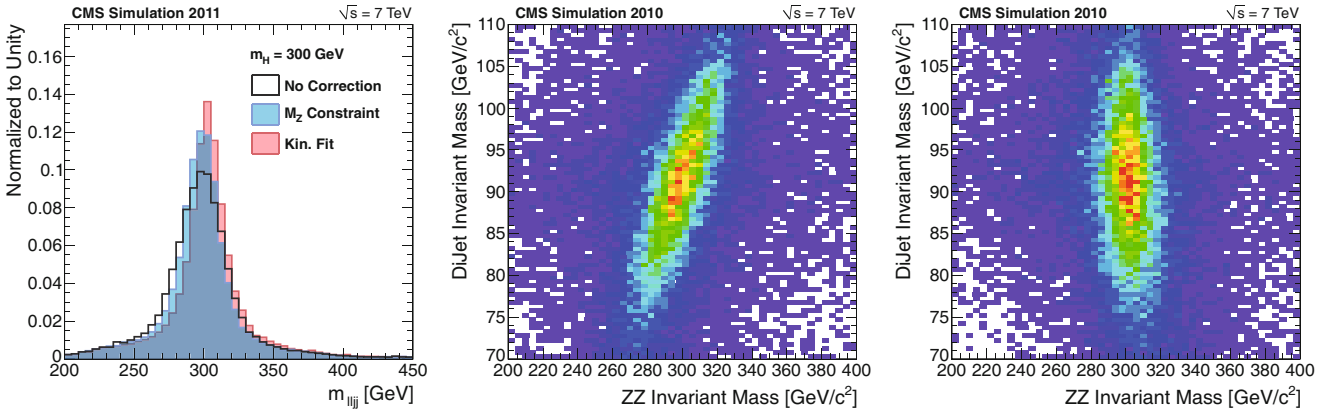

Fig. 9. Left: reconstructed Higgs invariant mass spectra in signal events with a 300 GeV Higgs boson. The black histogram respresents the uncorrected distribution, the blue histogram is obtained after imposing the Z boson mass to the dijet quadrimomentum, the red one by applying the kinematic fit. All distributions are normalized to unit area. Center and right: correlation between the reconstructed dijet invariant mass and the reconstructed diboson invariant mass in signal events with $m_H = 300\,\text{GeV}$, before (center) and after (right) the kinematic fit.

It was found that in the 2-tag category a significant contribution to the backgrounds after selections was originating from $t\bar{t}$ events. To contrast this source of background, a requirement is introduced on the significance of the reconstructed particle-flow missing transverse energy ($\text{PF}\not{E}_\text{T}$). This is done by defining a likelihood-ratio discriminant λ, which, through the knowledge of the expected resolutions on the event's reconstructed jets, compares the hypothesis that the event presents a true missing transverse energy ($\not{E}_\text{T}$) equal to the measured $\text{PF}\not{E}_\text{T}$, to the null hypothesis ($\not{E}_\text{T} = 0$). The observed distribution of $2\ln\lambda$ on $4.6\,\text{fb}^{-1}$ of data passing preselection requirements is shown in fig. 8 (right). The distribution for events coming from a 400 GeV Higgs boson decay, scaled by a factor 100, is overlaid, even though little to no dependance is observed as a function of the hypothetical signal mass. We therefore introduce an additional requirement, in the 2-tag category only, that the event $\text{PF}\not{E}_\text{T}$ satisfies the requirement $2\ln\lambda < 10$. This ensures high efficiency ($> 97\%$) on signal events, and is expected to reject more than 50% of the top background.

5.3 Kinematic fit to the decay chain

The aim of the analysis is to study the invariant mass spectrum of the dilepton+dijet system, in order to search for signal-like excesses. Signal events are resonant in this variable, as the decay of a massive particle is involved. If no biases are introduced at selection level, signal events will present an invariant mass peak centered at the Higgs boson mass. The significance of the excess depends on the width of the invariant mass peak, which will have two components: an intrinsic one, which depends on the Higgs intrinsic decay width, which can be very large for massive Higgs bosons; and the effect of detector resolutions, which is dominated by the resolution on jets.

In order to contrast the effect of jet resolutions on the invariant mass peak, an additional piece of information may be exploited: jets in signal events are known to stem from the decay of a Z boson, therefore their invariant mass should be compatible with the Z boson mass (m_Z). Hence imposing to the dijet system to have an invariant mass equal to m_Z is expected to improve the final invariant mass resolution for signal events, whereas no significant effect is expected to be introduced in the main backgrounds, which are non-resonant in the dijet system.

The simplest way of imposing the m_Z mass to the dijet system is that of rescaling the dijet quadrimomentum as a whole, modifying its energy in order to obtain the needed mass. This simple procedure already significantly improves the invariant mass scale and resolution of signal events, as can be seen in fig. 9 (left), where the uncorrected dilepton-dijet invariant mass spectrum (black) is compared to the one obtained by applying this rescaling (blue) for signal events with $m_H = 300\,\text{GeV}$. Though effective, this procedure is clearly suboptimal, as it treats both jets "democratically", without exploiting the prior knowledge we have on their expected resolutions. We know for instance that jets with higher energies are expected to be reconstructed with higher precision than jets with lower energies, as well as the fact that different detector regions have different expected jet reconstruction performance.

A more powerful approach, that makes use of the information on individual jets, is to perform a kinematic fit to the dijet system. The fit takes as input the quadrimomenta of the two jets, and makes use of the knowledge of the expected jet transverse momentum and position resolutions, as a function both of transverse momentum and pseudorapidity. It then proceeds in modifying the jet quadrimomenta, compatibly with the expected resolution, until the dijet system assumes the Z mass.

Table 2. Summary of selection requirements in the $H \to ZZ \to \ell^+\ell^-q\bar{q}$ analysis, split into the three analysis categories. The angular likelihood discriminant (aLD) requirement depends on the reconstructed Higgs boson candidate mass (m_H). The quark-gluon likelihood discriminant (QG LD) and PF$\not\!\!E_T$ significance ($2\ln\lambda$) are enforced respectively only in the 0 b-tag and the 2 b-tag categories.

0 b-tag	1 b-tag	2 b-tag
Lepton HLT/ID/Isolation and $p_T > 40/20\,\text{GeV}$		
Jet $p_T > 30\,\text{GeV}$ and $\lvert\eta\rvert < 2.4$		
$70 < m_{\ell\ell} < 110\,\text{GeV}$		
$75 < m_{jj} < 105\,\text{GeV}$		
Kinematic fit to the $Z \to q\bar{q}$ decay chain		
aLD $> 0.00025 \cdot m_H + 0.55$	aLD $> 0.000656 \cdot m_H + 0.302$	aLD > 0.5
QG LD > 0.1		$2\ln\lambda < 10$

Table 3. Expected yields of signal (signal efficiency is shown in parentheses) and background per fb^{-1} based on simulation in the 0 b-tag category. In each case the two numbers show $2e2j/2\mu2j$ expectations. For each considered signal mass (m_H), events are counted only in the $-6\%/+10\%$ window about the nominal Higgs mass.

m_H [GeV]	Signal	Z+jets	Diboson	$t\bar{t}/tW$	Total BG
250	2.2/2.5 (2.1%/2.3%)	81/99	2.8/3.2	0.92/0.97	85/105
300	2.4/2.5 (3.0%/3.1%)	40/53	1.7/2.4	0.25/0.36	42/55
350	2.5/2.5 (3.4%/3.5%)	21/28	1.3/1.5	0.11/0.1	23/29
400	1.8/1.8 (3.3%/3.3%)	11/15	0.74/0.84	0.0076/0.079	12/16
450	1.1/1.1 (2.9%/3.0%)	8.3/7.4	0.59/0.55	0.03/0.0067	8.9/8
500	0.61/0.67 (2.6%/2.9%)	3.3/3.7	0.32/0.4	0/0.0046	3.6/4.1

The kinematic fit further improves the resolution on the final reconstructed Higgs invariant mass peak, as can be seen in fig. 9 (red). For masses heavier than $\sim 400\,\text{GeV}$, little margin of improvement is expected, because the Higgs intrinsic width becomes the dominant factor in the determination of the invariant mass peak width.

An additional feature of the kinematic fit is that it removes the correlation between the reconstructed dijet and the diboson invariant masses. These two quantities are expected to be correlated because fluctuations in the measured jet momenta, driven by their relatively poor resolutions, will reflect with similar biases in both variables, as can be seen for a 300 GeV Higgs boson in fig. 9 (center). Once the kinematic fit is applied, the dependance of the diboson invariant mass on jet resolutions is minimized, hence the correlation is removed (right).

5.4 Summary of selection requirements

Table 2 summarizes the selection requirements for the $H \to ZZ \to \ell^+\ell^-q\bar{q}$ analysis. The main discrimination power is provided by the angular likelihood discriminant: events are required to satisfy a threshold which depends on the reconstructed Higgs invariant mass, as shown in the table. The dependance on the Higgs mass is different in the three categories. Additional selections are enforced in specific categories only: namely the quark-gluon discrimination requirement in the 0-tag category, and the PF$\not\!\!E_T$ significance ($2\ln\lambda$) requirement in the 2-tag category.

Tables 3, 4 and 5 instead show, respectively for the three b-tag categories, the expected yields and signal efficiencies per fb^{-1} of integrated luminosity in the $-6\%/+10\%$ m_{lljj} range about the nominal Higgs boson mass, for six hypothetical signal masses. Expected event yields in the electron and muon channel are quoted separately. Backgrounds are here shown broken up in their different contributions.

Table 4. Expected yields of signal (signal efficiency is shown in parentheses) and background per fb^{-1} based on simulation in the 1 b-tag category. In each case the two numbers show $2e2j/2\mu2j$ expectations. For each considered signal mass (m_{H}), events are counted only in the $-6\%/+10\%$ window about the nominal Higgs mass.

m_{H} [GeV]	Signal	Z+jets	Diboson	$t\bar{t}/tW$	Total BG
250	1.7/1.9 (1.6%/1.8%)	69/81	2.4/3	7.4/8.4	79/93
300	1.7/1.9 (2.1%/2.4%)	39/48	1.7/1.9	2.8/3.8	44/54
350	1.9/2.1 (2.6%/2.83%)	23/30	0.91/1.2	1/0.93	25/32
400	1.5/1.6 (2.6%/2.8%)	14/19	0.71/0.75	0.34/0.23	15/20
450	0.93/0.98 (2.6%/2.7%)	11/11	0.43/0.5	0.18/0.026	12/11
500	0.55/0.58 (2.4%/2.5%)	7.6/5.6	0.36/0.49	0.065/0.051	8/6.1

Table 5. Expected yields of signal (signal efficiency is shown in parentheses) and background per fb^{-1} based on simulation in the 2 b-tag category. In each case the two numbers show $2e2j/2\mu2j$ expectations. For each considered signal mass (m_{H}), events are counted only in the $-6\%/+10\%$ window about the nominal Higgs mass.

m_{H} [GeV]	Signal	Z+jets	Diboson	$t\bar{t}/tW$	Total BG
250	0.71/0.79 (0.66%/0.74%)	5.3/4.8	0.41/0.35	1.2/1.2	6.8/6.3
300	0.82/0.8 (1.0%/1.0%)	2.7/3.1	0.26/0.33	0.48/0.76	3.4/4.2
350	0.9/0.95 (1.2%/1.3%)	1.3/1.5	0.2/0.22	0.14/0.19	1.7/1.9
400	0.7/0.74 (1.3%/1.3%)	0.45/1.3	0.1/0.16	0.022/0.0084	0.58/1.5
450	0.46/0.49 (1.3%/1.3%)	0.63/1.3	0.097/0.16	0.0042/0.048	0.73/1.5
500	0.29/0.3 (1.3%/1.3%)	0.87/0.8	0.1/0.089	0/0.062	0.97/0.95

6 Background estimation

As the adopted event selection does not depend in any way on the hypothetical Higgs boson mass, but rather on the reconstructed dilepton-dijet invariant mass m_{lljj}, after the final selection is applied to the data we have a total six m_{lljj} distributions, one per b-tag category $(0, 1, 2)$ times one per lepton flavour (e, μ). We analyze these distributions for different hypothetical Higgs boson signals, as the selection is expected to yield different efficiencies for different hypothetical Higgs masses. The distribution of background events, though, is unique in each of the six channels.

We do not intend to fully rely on the simulation to estimate the expected background yields after applying the event selection, therefore we measure the background directly from the data. This is done by analyzing the dijet invariant mass (m_{jj}) in an extended range, and splitting events in two separate regions:

- events which pass the nominal selection ($75 < m_{jj} < 105\,\mathrm{GeV}$) are placed in the *signal region*, and are of interest for the final analysis results;
- events which fail the analysis selection because of the value of m_{jj} are kept if they lie in the broader invariant mass interval of $60 < m_{jj} < 130\,\mathrm{GeV}$, and define the *sideband region*.

The thresholds which define the sideband region are the result of a compromise: they are tight enough to ensure that the kinematics of sideband events is similar to the ones in the signal region, and wide enough so that the available amount of data is comparable in the two regions. As Higgs events present the hadronic decay of a Z boson, the sideband region is reasonably depleted of signal. On the other hand, most of the backgrounds are not resonant in the dijet invariant mass variable (the only exception is the direct Z pair production), and are therefore expected to populate the signal and sideband region in similar fashion.

Even if the event kinematics, and therefore the resulting m_{lljj} distributions, are similar between the signal and sideband regions, they are not identical. In order to use sideband events to estimate the background yield in the signal region, the former have to be corrected to take into account this difference. This is done by accessing the Monte Carlo simulation: for each b-tag category, the shapes of the m_{lljj} distributions in the signal and sideband regions are

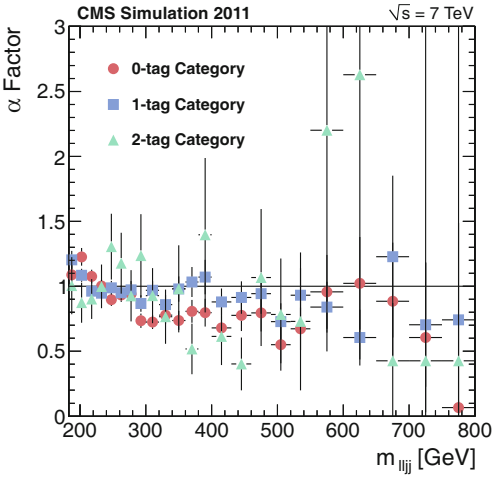

Fig. 10. Sideband region correction factor (α) as a function of the reconstructed dilepton-dijet invariant mass (m_{lljj}) in the three b-tag categories: 0-tag (red circles), 1-tag (blue squares), 2-tag (green triangles).

compared, and a bin-to-bin ratio $\alpha(m_{lljj})$ is computed. The value of $\alpha(m_{lljj})$, in the three b-tag categories, is shown in fig. 10: it is not very different from unity, therefore the entity of the correction (and of the possible uncertainty it implies) is small.

The comptutation of the α ratio enables us to estimate the number of background events $N_{\rm bkg}$ at a given m_{lljj} invariant mass. This is done by taking the number of observed events in the data sidebands ($N_{\rm sb}$) and correcting it with the following formula:

$$N_{\rm bkg}(m_{lljj}) = N_{\rm sb}(m_{lljj}) \times \frac{N_{\rm bkg}^{\rm MC}(m_{lljj})}{N_{\rm sb}^{\rm MC}(m_{lljj})} \equiv N_{\rm sb}(m_{lljj}) \times \alpha(m_{lljj}),$$

where the corresponding Monte Carlo yields are indicated with a superscript.

The resulting α-corrected m_{lljj} sideband distribution constitutes our data-driven estimate of the signal region background yield. In order to minimize the effect of statistical fluctuations originating from the limited amount of data, the distribution is fitted with an empirical functional form, which was found to successfully describe the shape obtained on the simulation: the product of a Fermi-Dirac, for the steep low-mass turn-on, and a Crystal-Ball function, for the kinematical peak around 200 GeV and the high-mass tail.

The function has a total of six floating parameters: two from the Fermi-Dirac function (the equivalent temperature and the position of the transition), and four from the Crystal-Ball (the mean and width of the Gaussian, the Gaussian/power-law transition position, and the exponent of the power-law) . The function is used in an unbinned, maximum-likelihood fit to the sideband distribution in the simulation, with all parameters free to vary, taking advantage of the high number of available Monte Carlo events. It is then fitted to the α-corrected sideband distribution observed in the data, but only two Crystal-Ball parameters are kept floating in the fit procedure (the Gaussian width and the power-law exponent), whereas all other parameters are fixed to the values obtained on the simulation.

The results of the unbinned, maximum-likelihood fits to the α-corrected m_{lljj} sideband distribution are shown in fig. 11: 0 b-tag category on the left, 1 b-tag in the center, and 2 b-tag on the right. Plots on the top (bottom) row are in linear (logarithmic) scale. The result of the fit is shown with a blue curve, and 68% (95%) fit uncertainty bands are shown with a green (yellow) shade. These represent the background estimate for the signal region events in the three b-tag categories.

7 Systematic uncertainties

The background estimation in this analysis, as has been shown, is obtained directly from the data, by analysing the dijet invariant mass sidebands. We here treat the effects which could affect signal efficiency. They are summarized in table 6 and detailed in the following.

Lepton reconstruction. Systematic uncertainties originating from lepton trigger, reconstruction and identification have been obtained directly on data, by measuring the relative efficiencies with a tag-and-probe method [29] applied to leptons originating from the decay of a Z boson. This translates into a signal efficiency uncertainty of 2.7% for muons and 4.5% for electrons.

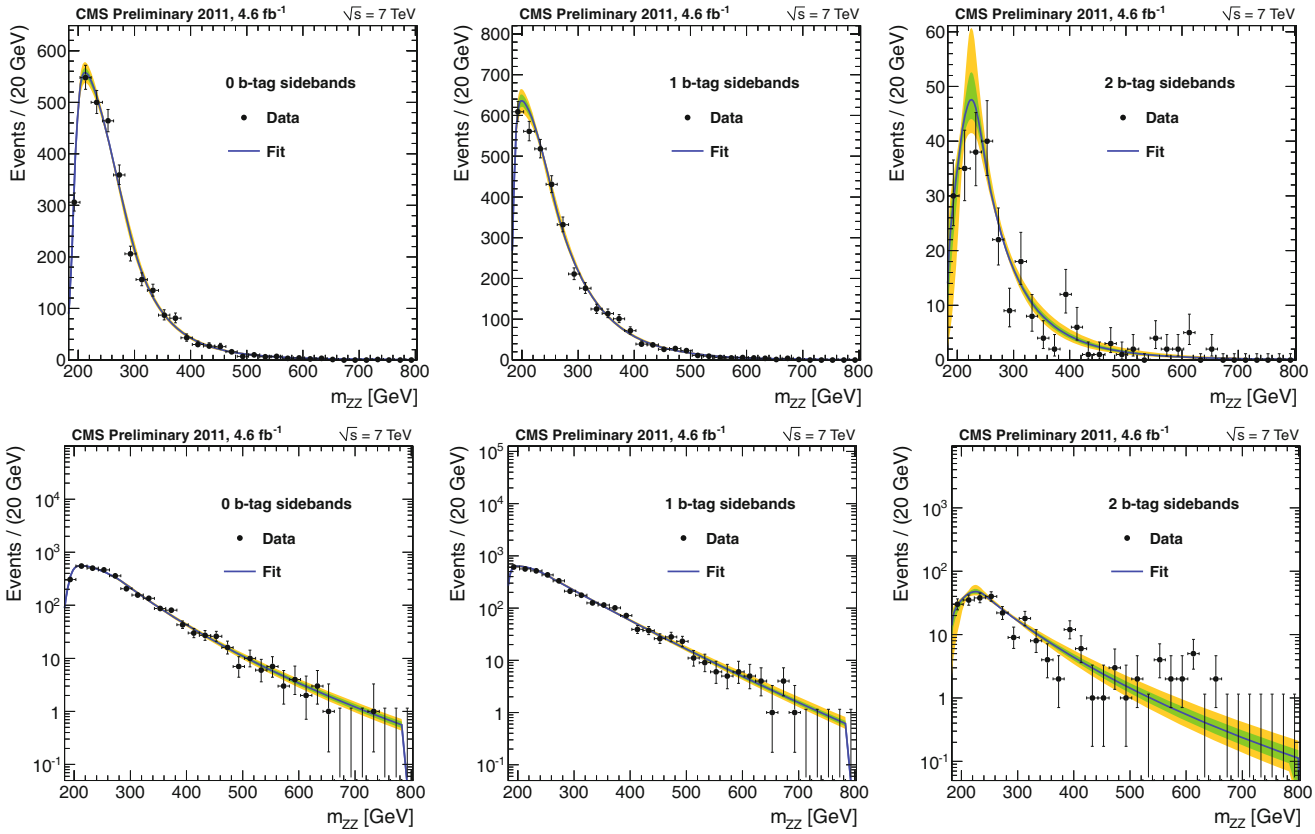

Fig. 11. Results of the unbinned, maximum-likelihood fits to the alpha-corrected data sidebands: 0 b-tag category on the left, 1 b-tag in the middle, and 2 b-tag on the right. Plots on the top (bottom) column are in linear (logarithmic) scale. The result of the fit is shown with a blue curve, and 68% (95%) fit uncertainty bands are shown with a green (yellow) shade.

Table 6. Summary of systematic uncertainties on signal efficiency, separated by source. Uncertaintes common to all three b-tag categories are placed in the center column, whereas sources which have different effects on the three categories are reported with distinct contributions. See text for details.

Source	0 b-tag	1 b-tag	2 b-tag
Muon reconstruction		2.7%	
Electron reconstruction		4.5%	
Jet energy scale and resolution		1–5%	
Pile up		4%	
b-tagging	3%	1%	20%
Quark-gluon discrimination	4.6%	–	–
Missing E_{T}	–	–	3%
Higgs cross section		13–18%	
Higgs production (PDF)		3%	
Higgs production (HQT)	2%	5%	3%
Higgs production (VBF)		1%	
Higgs boson width		1–30%	
Luminosity		4.5%	

Jet energy scale. Jet transverse momenta have been varied within the measured jet energy scale uncertainties, and the difference in signal yield taken as an estimate of the jet energy scale uncertainty. To evaluate the systematic uncertainty due to jet resolutions an additional Gaussian smearing factor was introduced to simulate a worse jet transverse momentum resolution. Depending on the signal hypothetical mass, uncertainties between 1 and 5% were obtained.

Pile-up. The simulated signal samples have been divided in two subsets, depending on whether events presented a number of reconstructed primary vertices in the event superior or inferior to seven, which is the approximate position of the mode of the data. The full event selection was then applied to the two subsets and the maximal difference in signal efficiency with respect to the inclusive sample was taken as uncertainty. It was found to be equal to 4%.

b-tagging. We have taken as a systematic uncertainty relative to this method the observed difference in signal yield when varying the b-tagging scale factors by on standard deviation. This is found to amount to 3%, 1% and 20%, respectively for the 0-tag, 1-tag and 2-tag categories. It must be noted that these variations are correlated, as, for instance, a decrease in the 2-tag category yield will imply an increase in the 1-tag yield.

Quark-gluon discrimination. A light-quark enriched control sample is identified by photon+jet events, in which the leading jet originates from light quarks in more than 90% of the cases. To eliminate possible contributions from bottom quarks, a b-tag veto is enforced. The data-MC differences in the quark-gluon tagger are studied as a function of the photon transverse momentum. No significant deviation is found, therefore the statistical uncertainty of the efficiency computed on the data control sample is taken as a systematic uncertainty: it amounts to 4.6%.

Missing transverse energy. Missing transverse energy affects directly only the 2 b-tag category. The dominant effects which could concur in generating uncertainty derive from the knowledge of the rest of the event, such as jet energy reconstruction and pile-up. Therefore, most of the related uncertainty should be covered by the corresponding systematic uncertainties. The adopted requirement on missing transverse energy significance is very loose on signal events, and translates in a maximal inefficiency of 3%. We postulate that the resulting uncertainty does not surpass this value.

Signal cross section. We follow the recommendation found in [30] to vary the Higgs cross section within its expected uncertainties at each mass point. The total uncertainty, weighed on the different production processes, is in the range (13.4–18.0)%. We note that this uncertainty is relevant only for the measurement of the ratio to the SM expectation, while it does not affect the absolute cross section measurement.

Proton parton density functions. The uncertainty related on the proton parton distribution functions (PDFs) is evaluated following the PDF4LHC [31] recommendations. This is done by evaluating the selection efficiencies (and the relative error sets) of three different sets of PDFs: CT10 [32], MSTW2008NLO [33], NNPDF2.1 [34]. The corresponding uncertainty on signal efficiency is then taken as the envelope of the three error bands, and translates into a 2–4% effect, with a dependance on the Higgs mass and on the b-tag category.

Higher order contributions. Missing higher orders in perturbation theory, which are not included in the POWHEG NLO computation, may modify the Higgs production kinematics, and therefore affect the selection efficiency. The related uncertainty was quantified through the use of the HQT [35] program, which includes NNLL effects. The POWHEG sample is reweighed in order to match the Higgs transverse-momentum spectrum predicted by HQT, and the corresponding deviation from the nominal signal efficiency is taken as uncertainty. The effect was maximal for $m_H = 200\,\text{GeV}$, where the reweighing translated into an efficiency drop of 2%, 5%, 3%, respectively for the 0-, 1-, and 2-tag categories. At higher masses, `POWHEG` and HQT are found to be in better agreement, and the deviation is found to be within 1%. Conservatively, the observed effect at $m_H = 200\,\text{GeV}$ was taken as systematic uncertainty for all masses.

VBF efficiency. Only the contribution of gluon fusion was considered during the tuning of the analysis and the interpretation of the results, as it contributes to about 90% of the total cross section over most of the mass range. A real signal, though, would contain the correct mixture of all the production processes, and therefore the Vector Boson Fusion (VBF) channel, which has in general different final state kinematics, may modify the selection efficiency on signal. We evaluated the corresponding uncertainty as the difference in acceptance between the two production processes, and multiplied it by the expected VBF fractional contribution to the total cross section. The results of this procedure identify an uncertainty of about 1%.

Higgs width. In this study the cross section for on-shell Higgs production and decay was made in the zero-width approximation, and acceptance estimates are obtained with Monte Carlo simulations that are based on *ad-hoc* Breit-Wigner distributions for describing the Higgs boson propagation. Recent analyses show that the use of a QFT-consistent Higgs propagator, allowing also for the off-shellness of the Higgs boson, dynamical QCD scales and interference effects between Higgs signal and backgrounds will result, at Higgs masses above 300 GeV, in a sizable effect on conventionally defined but theoretically consistent parameters (mass and width) that describe the propagation of an unstable Higgs boson [36,30,37]. These effects are estimated to amount to an additional uncertainty (U) on the theoretical cross section which depends on the Higgs boson mass (m_H), and we evaluate it using the following formula:

$$U(m_\text{H}) = 150\% \cdot (m_\text{H}\,[\text{TeV}])^3.$$

As can be seen this uncertainty is neglibigle for masses inferior to 300 GeV, but grows rapidly with mass.

LHC luminosity. The uncertainty on the luminosity measurement is 4.5% [38].

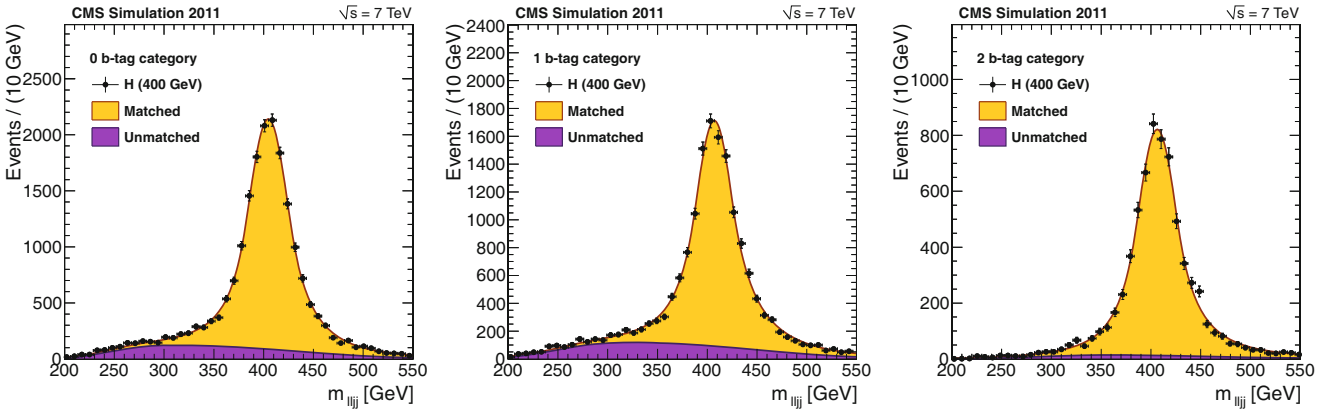

Fig. 12. Distribution of the reconstructed dilepton-dijet invariant mass (m_{lljj}) for a 400 GeV Higgs boson signal and in the three b-tag categories (0 b-tag on the left, 1 b-tag in the center and 2 b-tag on the right). The spectra obtained on the simulation are shown with black markers, and the results of the fit is shown which a continuous line. Contributions from matched (yellow) and unmatched (violet) events are shown separately.

8 Statistical interpretation of results

The strategy adopted by the CMS collaboration is to search for the Standard Model Higgs boson in a total of 173 mass points across the 114–600 GeV invariant mass range. This analysis has limited statistical power for masses below the ZZ production threshold, therefore we will focus on the high-mass range 200–600 GeV, which comprises of a total of 73 mass points. In this section we describe how we model the presence of a hypothetical Higgs signal, and the statistical methods adopted to convert the analysis outcome into a statement on the Higgs boson's existence.

8.1 Modeling of the signal

In each of the six analysis categories, the same m_{lljj} invariant mass spectrum is analyzed for numerous hypothetical Higgs boson signals, varying its postulated mass. Because of computing limitations, though, we are not able to generate Monte Carlo samples at each mass point in which we intend to perform a search. Rather, samples equivalent to high integrated luminosities have been generated at a number of pivotal mass points, where the behaviour of the expected signal is studied, and results are interpolated at every intermediate mass point.

Two quantities need to be parametrized as a function of the Higgs boson mass: the selection efficiency, and the shape of the expected signal. The selection efficiency is computed at the specific hypothetical mass points where simulated samples have been generated, and these points are subsequently fitted, separately in the electron and the muon channels, with polynomial functions.

The modeling of the signal shape is done by subdividing signal events which pass the analysis selection in two categories: those which have jets which are correctly matched to the quarks originated in the Higgs boson decay ("matched" events), and those in which an incorrect jet pair has been chosen by the selection algorithm, because of signal self-combinatorics ("unmatched" events). This is done by accessing the generator information in signal samples, and performing a matching between the reconstructed jets and the generator quarks produced in the Higgs decay. The reconstructed invariant mass distribution of matched events is parametrized with a double Crystal-Ball function, in order to take into account detector resolutions. Unmatched events are described with a triangle function smeared with a Crystal-Ball, which was empirically found to adeguately describe the shape observed in the simulation. The sum of these two functions defines the adopted parametrization of the shape of the invariant mass distribution of signal events.

In order to have a signal parametrization valid for any given Higgs mass, unbinned maximimum-likelihood fits are performed to the invariant mass spectra obtained on the simulation, for all the available mass points and separately in the three b-tag categories. Examples of the results of such fits are shown in fig. 12, for a 400 GeV Higgs boson signal, in the three b-tag categories (0 b-tag on the left, 1 b-tag in the center and 2 b-tag on the right). An overall good agreement between the fit results and the shape of the spectra is observed. Once the fits are performed at all of the mass points made available by the Monte Carlo production, the values of the fit parameters are studied as a function of the Higgs mass (m_{H}), and are fitted with linear or quadratic functions, so as to obtain a smooth dependance on m_{H}.

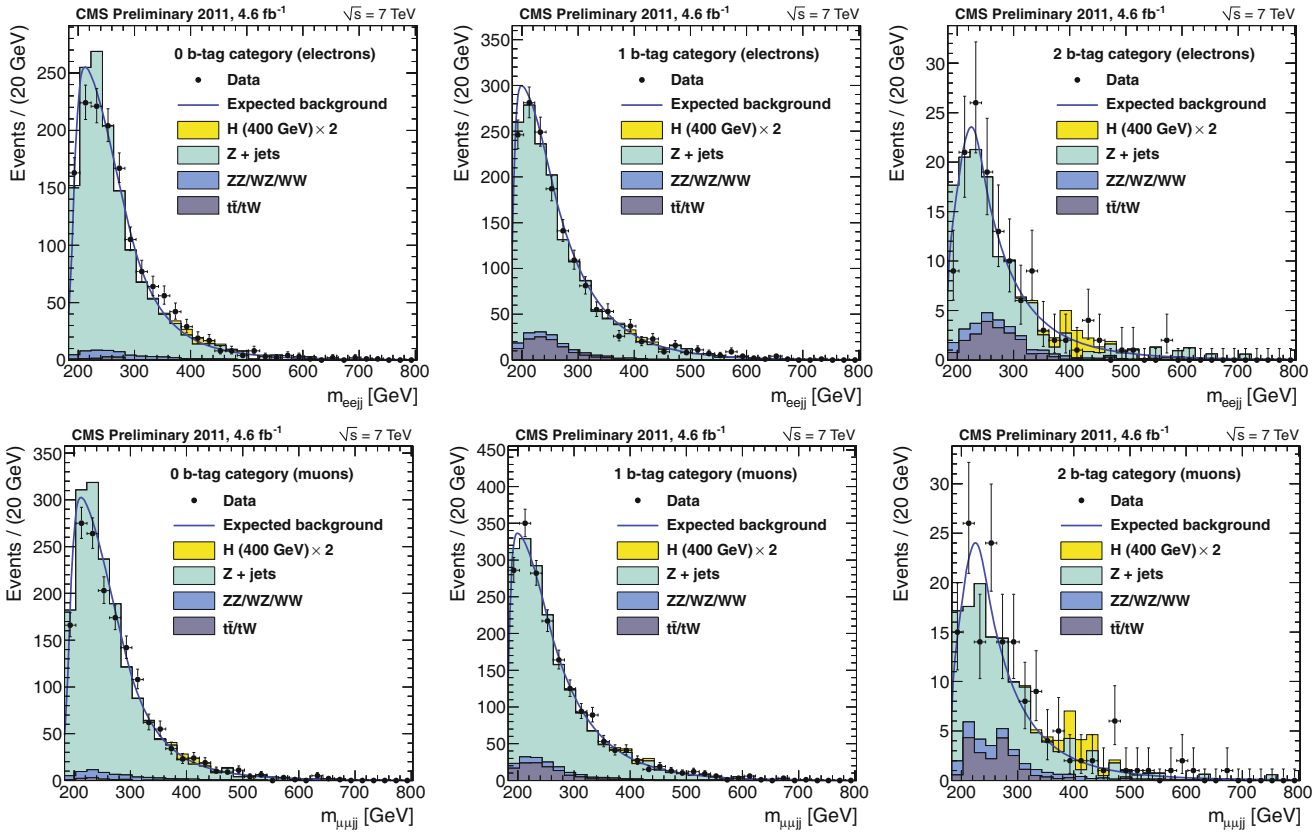

Fig. 13. The dilepton-dijet invariant mass after full selection in the six analysis categories: 0 b-tag on the left, 1 b-tag in the center, 2 b-tag on the right. The contributions of the electronic (top) and muonic (bottom) channels are shown separately. The data-driven estimate of the background contribution is overlaid as a blue line. The expected yield, in the simulation, of the dominant backgrounds, as well as of a 400 GeV signal enhanced by 2 (yellow), is shown as a comparison.

8.2 Statistical analysis

The observed dilepton-dijet invariant mass spectra on $4.6\,\mathrm{fb}^{-1}$ of 2011 data, after the full selection is applied in the six analysis categories, are shown in fig. 13: 0 b-tag to the left, 1 b-tag in the center, 2 b-tag to the right. The contributions of the electronic (top) and muonic (bottom) channels are shown separately. The data-driven estimate of the background contribution, as extrapolated from the dijet invariant mass sideband region events (as described in sect. 6), is overlaid as a blue line. The expected yield, in the simulation, of the dominant backgrounds, as well as of a 400 GeV signal enhanced by 2, is shown as a comparison.

For each mass hypothesis, we perform a simultaneous likelihood fit of the six m_{lljj} distributions using the statistical approaches discussed in [39]. As no significant excess over the background prediction is observed, we proceed to set limits on the Standard Model Higgs production. The method we adopt for reporting limits is the $\mathrm{CL_s}$ modified frequentist technique [40]. All results are validated by using two independent sets of software tools, the RooStats package [41] and L&S [42].

Based on the expected normalization and shape of the m_{lljj} distribution, for signal and background, and the corresponding systematic uncertainties, we generate a large number of random pseudo-experiments. For each of them, the expected background distribution is generated, a likelihood fit is performed, and an exclusion limit is extracted. The median of the results is taken as central value of the expected statistical power of the analysis, and the distribution is integrated to define 68% and 95% probability intervals about the median. These values are then compared to the observed limit, which is obtained by the fit to the analyzed data.

Observed (markers) and expected (dashed line) exclusion limits on the product of the Higgs boson production cross section and the branching fraction of $\mathrm{H} \to \mathrm{ZZ}$ are presented in fig. 14 (left) using the $\mathrm{CL_s}$ technique. The expected limit also shows the 68% and 95% probability ranges, respectively marked by a green and a yellow shade. As a comparison, the expectation of the production cross section times branching fraction are shown for the Standard Model (SM), and for an extensions of the latter (SM4), in which a fourth generation of massive fermions is introduced [43,44, 37]. The main difference from the SM Higgs production is that, due to the couplings introduced by the additional fermions, the signal production cross section is enhanced by a factor which varies between 8.3 and 4.8 for a Higgs

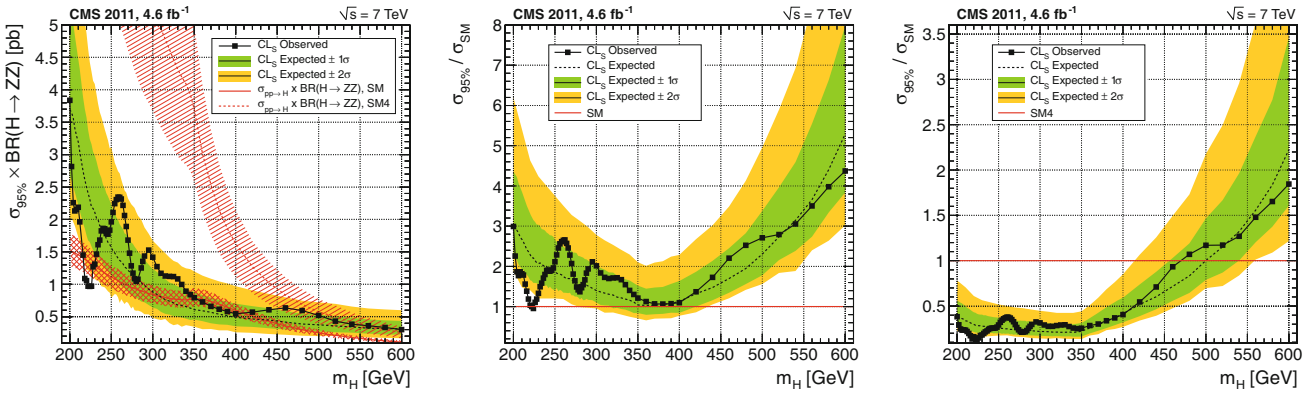

Fig. 14. Observed (markers) and expected (dashed line) 95% confidence level upper limit on the product of the Higgs boson production cross section (left), on the ratio of the production cross section to the SM one (center) and to the expectation of the SM4 model. The 68% and 95% ranges of expectation are also shown with green and yellow bands. The expected product of the SM Higgs production cross section and the branching fraction is shown as a red solid curve with a band indicating theoretical uncertainties at 68%. The same expectation in the SM4 model are shown with the upper red curve, with a band indicating theoretical uncertainties.

boson in the 200–600 GeV mass range. We assume the main uncertainties on the SM4 Higgs production cross section to be the same as for the gluon-fusion mechanism in the Standard Model but with an additional 10% uncertainty due to the electroweak radiative corrections. This additional uncertainty is added linearly to uncertainties from QCD renormalization and factorization scales, PDFs, and α_s.

We further incorporate uncertainties on the Higgs production cross section and present a limit on the ratio of the SM Higgs boson production cross section to the SM expectation in fig. 14 (center): the observed limit (markers) is compared to the expected one (dashed line), and the latter is provided of 68% (green) and 95% (yellow) probability bands. This search alone, with 4.6 fb^{-1} of data, reaches the sensitivity for a 95% confidence level exclusion of a Standard Model Higgs boson in two mass ranges: 224–226 and 360–400 GeV. As can be seen the observed exclusion presents a good degree of compatibility with expectations. The significant deviation from the expected trend observed around 225 GeV been deeply scrutinized, and was found compatible with a statistical effect, driven by the observed under-fluctuation in both the electron and muon 0 b-tag channels.

A similar limit on the ratio to the Higgs boson production cross section in the SM4 model is shown in fig. 14 (right). A range of SM4 Higgs mass hypotheses are excluded between 200 and 460 GeV at 95% confidence level.

9 Conclusions

We have presented a search for a massive Higgs boson in the decay channel $\text{H} \to \text{ZZ} \to \ell^+\ell^-\text{q}\bar{\text{q}}$. This channel presents jets in the final state, which threaten to degrade the analysis performance both by worsening its mass resolution and by increasing the possible sources of backgrounds. Therefore stringent requirements are imposed on the jet reconstruction performance, as both an accurate calibration and a good resolution on the measurement of their quadrimomenta are needed. This is achieved by utilizing particle-flow jet reconstruction, calibrated *in situ* with photon+jet events. The resolution on the Higgs invariant mass is further boosted by the application of a kinematic fit to the hadronic decay of the Z boson.

The analysis selection maximises the sensitivity to a presence of a Higgs boson signal by pursuing two main directives:

– an angular analysis, to discriminate events compatible with the decay of a scalar boson from non-resonant backgrounds;

– the use of jet parton flavour tagging as means of background rejection and sensitivity maximization.

After applying the full selection on 4.6 fb^{-1} of data collected in 2011 by the CMS detector, no evidence for the presence of a Standard Model Higgs boson has been found, and we set upper limits on its production cross section, reaching sensitivity to the Standard Model prediction in two mass ranges: 224–226 and 360–400 GeV. We also constrain the presence of a Higgs boson in the context of an extended Standard Model, in which a fourth generation of massive fermions is introduced, by excluding it in the 200–460 GeV mass range, at 95% confidence level. When combined to the other searches performed at CMS, the Standard Model Higgs boson is excluded in a broad mass range: between 127 and 600 GeV.

References

1. Peter W. Higgs, Phys. Lett. **12**, 132 (1964).
2. Peter W. Higgs, Phys. Rev. Lett. **13**, 508 (1964).
3. F. Englert, R. Brout, Phys. Rev. Lett. **13**, 321 (1964).
4. G.S. Guralnik, C.R. Hagen, T.W.B. Kibble, Phys. Rev. Lett. **13**, 585 (1964).
5. L. Evans, P. Bryant (Editors), JINST **3**, S08001 (2008).
6. CMS Collaboration, JINST **3**, S08004 (2008).
7. ATLAS Collaboration, JINST **3**, S08003 (2008).
8. CMS Collaboration, Phys. Lett. B **716**, 30 (2012).
9. ATLAS Collaboration, Phys. Lett. B **716**, 1 (2012).
10. CMS Collaboration, JHEP **12**, 1 (2012).
11. S. Chatrchyan *et al.*, JINST **3**, S08004 (2008).
12. Serguei Chatrchyan *et al.*, JINST **7**, P10002 (2012).
13. CMS Collaboration, *Electron reconstruction and identification at* $\sqrt{s} = 7\,\text{TeV}$, in *CMS Physics Analysis Summary* CMS-PAS-EGM-10-004, 2010.
14. V. Khachatryan *et al.*, JHEP **01**, 080 (2011).
15. CMS Collaboration, *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and* E_T^{miss}, in *CMS Physics Analysis Summary* CMS-PAS-PFT-09-001, 2009.
16. Matteo Cacciari, Gavin P. Salam, Gregory Soyez, JHEP **04**, 063 (2008).
17. CMS Collaboration, JINST **6**, P11002 (2011).
18. Matteo Cacciari, Gavin P. Salam, Gregory Soyez, JHEP **04**, 005 (2008).
19. CMS Collaboration, *Identification of b-quark jets with the CMS experiment*, 2012.
20. Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, Tim Stelzer, JHEP **06**, 128 (2011).
21. T. Sjöstrand, S. Mrenna, P.Z. Skands, JHEP **05**, 026 (2006).
22. P. Nason, JHEP **11**, 040 (2004).
23. S. Frixione, P. Nason, C. Oleari, JHEP **11**, 070 (2007).
24. S. Alioli, P. Nason, C. Oleari, E. Re, JHEP **07**, 06 (2008).
25. S. Agostinelli *et al.*, Nucl. Instrum. Methods A **506**, 250 (2003).
26. B. Abbott *et al.*, Nucl. Instrum. Methods Phys. Res. A **424**, 352 (1999).
27. G. Alexander *et al.*, Z. Phys. C Part. Fields **69**, 543 (1996) DOI: 10.1007/s002880050059.
28. Yanyan Gao, Andrei V. Gritsan, Zijin Guo, Kirill Melnikov, Markus Schulze, Nhan V. Tran, Phys. Rev. D **81**, 075022 (2010).
29. V. Khachatryan *et al.*, JHEP **11**, 1 (2011) DOI: 10.1007/JHEP01(2001)080.
30. LHC Higgs Cross Section Working Group, S. Dittmaier, C. Mariotti, G. Passarino, R. Tanaka (Editors), *Handbook of LHC Higgs Cross Sections:* 1. *Inclusive Observables*, CERN-2011-002 (CERN, Geneva, 2011).
31. Sergey Alekhin *et al.*, The PDF4LHC Working Group Interim Report, 2011.
32. Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, Jon Pumplin, C.P. Yuan, New parton distributions for collider physics, 2010.
33. A.D. Martin, W.J. Stirling, R.S. Thorne, G. Watt, Eur. Phys. J. C **63**, 189 (2009) arXiv:0901.0002 [hep-ph] and `http://projects.hepforge.org/mstwpdf/`.
34. Richard D. Ball, Valerio Bertone, Francesco Cerutti, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Jose I. Latorre, Juan Rojo, Maria Ubiali, *Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology*, 2011.
35. G. Bozzi, S. Catani, D. de Florian, M. Grazzini, Nucl. Phys. B **737**, 73 (2006).
36. Giampiero Passarino, Christian Sturm, Sandro Uccirati, Nucl. Phys. B **834**, 77 (2010).
37. Charalampos Anastasiou, Stephan Buhler, Franz Herzog, Achilleas Lazopoulos, *Total cross-section for Higgs boson hadroproduction with anomalous Standard Model interactions* (2011).
38. Andrea Achilli, Rohit Hegde, Rohini M Godbole, Agnes Grau, Giulia Pancheri, Yogi Srivastava, Phys. Lett. B **659**, 137 (2007).
39. CMS Collaboration, *SM Higgs Combination*, in *CMS Physics Analysis Summary*, CMS-PAS-HIG-11-011, 2011.
40. A.L. Read, J. Phys. G **28**, 2693 (2002).
41. Lorenzo Moneta, Kevin Belasco, Kyle Cranmer, Alfio Lazzaro, Danilo Piparo, Gregory Schott, Wouter Verkerke, Matthias Wolf, *The RooStats Project*, September 2010.
42. M. Chen, A. Korytov, *Limits and significance*, `https://mschen.web.cern.ch/mschen/LandS/`.
43. N. Becerici Schmidt, S.A. Cetin, S. Istin, S. Sultansoy, Eur. Phys. J. C **66**, 119 (2010).
44. Qiang Li, Michael Spira, Jun Gao, Chong Sheng Li, Phys. Rev. D **83**, 094018 (2011).